Audio-Visual Speech Recognition in MISP2021 Challenge: Dataset Release and Deep Analysis

Hang Chen1, Jun Du1,*, Yusheng Dai1, Chin-Hui Lee2, Sabato Marco Siniscalchi2,4, Shinji Watanabe3, Odette Scharenborg6, Jingdong Chen7, Bao-Cai Yin5, Jia Pan5

1 University of Science and Technology of China, China 2 Georgia Institute of Technology, USA 3 Carnegie Mellon University, USA 4 Kore University of Enna, Italy 5 iFlytek, China 6 Delft University of Technology, The Netherlands 7 Northwestern Polytechnical University, China

Abstract

In this paper, we present the updated Audio-Visual Speech Recognition (AVSR) corpus of MISP2021 challenge, a large-scale audio-visual Chinese conversational corpus consisting of 141h audio and video data collected by far/middle/near microphones and far/middle cameras in 34 real-home TV rooms. To our best knowledge, our corpus is the first distant multi-microphone conversational Chinese audio-visual corpus and the first large vocabulary continuous Chinese lip-reading dataset in the adverse home-tv scenario. Moreover, we make a deep analysis of the corpus and conduct a comprehensive ablation study of all audio and video data in the audio-only/video-only/audio-visual systems. Error analysis shows video modality supplement acoustic information degraded by noise to reduce deletion errors and provide discriminative information in overlapping speech to reduce substitution errors. Finally, we also design a set of experiments such as frontend, data augmentation and end-to-end models for providing the direction of potential future work. The corpus1 and the code2 are released to promote the research not only in speech area but also for the computer vision area and cross-disciplinary research.

Index Terms: Audio-visual, speech recognition, speech enhancement, data augmentation

1. Introduction

Modern automatic speech recognition (ASR) systems still suffer from performance degradations in real world application scenarios (e.g., home and meeting). Even with advances in technology [1, 2, 3] and the publicity of large-scale corpora recorded in real environments [4, 5, 6], state-of-the-art robust ASR system still runs into performance plateaus, e.g., the CHiME-6 [7] dinner party scenario reaches a word error rate of about 40%, which falls short of the deploy ability of the application. Many researches [8, 9, 10] have shown visual cues can help speech perception, especially in noisy environments. Inspired by these discoveries, Audio-Visual Speech Recognition (AVSR) system has been developed.


Meanwhile, various audio-visual corpora were released for training. Conventional audio-visual corpora were collected in the controlled environment, such as GRID [25], OuluVS [26], OuluVS2 [27] and TCD-TIMIT [28], etc. These corpora are limited in duration, number of speakers and vocabulary size. More recently, an effort of the research community has been put to gather data from different sources in the wild, e.g., public media websites or TV broadcasts. One of the first audio-visual in-the-wild corpora is LRW [29]. The trend of collecting larger datasets has continued in subsequent collections which consist of materials from British television programs [15, 14], TED talks [30], [31], etc. These large-scale datasets have a vast variety of speakers, sentences, languages and visual/auditory environments, but lack sample-level information. All these corpora are in English, a few corpora [32, 33, 34] have been released to AVSR in Chinese, which is the most widely used language worldwide. Nevertheless, there is still a lack of a large-scale public audio-visual speech corpus recorded in real-world application scenarios, especially for Chinese.

MISP2021 challenge [35] presented a distant multi-microphones conversational audio-visual corpus recorded in the home TV scenario, where several people are chatting in Chinese while watching TV and interacting with a smart speaker/TV in a living room. During the challenge, the corpus was only released to the registered participants and lacked a detailed corpus analysis. In this work, we have resolved authorization and storage issues to fully release the updated AVSR corpus of MISP2021 Challenge to all researchers and make a dataset update including correcting the asynchronous sample in the training/development set and adding more data to increase the data diversity of the evaluation set. Moreover, we provide extensive experiments for various baselines and deep analysis of the corpus. Specifically, we design a set of experiments of the audio-only/video-only/audio-visual speech recognition systems with all far/middle/near audio data and far/middle video data. Error analysis is conducted to explain that video modality supplement acoustic information to reduce deletion errors in noise conditions and provide discriminative information to reduce substitution errors in speech overlap. Finally, we also conduct a set of experiments such as frontend, data augmentation and end-to-end models to provide the direction of future work.

In the following, Section 2 introduces the updated corpus.
Figure 1: Illustration of an overall framework for hybrid ASR/VSR/AVSR systems with far/middle/near audio and far/middle video.

Table 1: Overview of the updated MISP2021-AVSR corpus.

<table>
<thead>
<tr>
<th>ConfigID</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>1V</td>
<td>Off</td>
<td>On</td>
<td>Off</td>
</tr>
<tr>
<td>Time</td>
<td>Day</td>
<td>Night</td>
<td>Day</td>
<td>Night</td>
</tr>
<tr>
<td>Light</td>
<td>on</td>
<td>off</td>
<td>on</td>
<td>off</td>
</tr>
</tbody>
</table>

Table 2: Scene configurations.

We describe the overall framework of systems in Section 3. A deep analysis combining experimental results and examples is presented in Section 4. We conclude in Section 5.

2. Dataset

As shown in Table 1, the updated MISP2021-AVSR corpus contains 141.24 hours of audio and video data from 253 native Chinese (98 males and 165 females) speaking Mandarin without strong accents which collected in 34 real real-home TV rooms. There is no overlap in speakers and recording rooms among the data in each subset. There are multiple microphone arrays and cameras used to collect far/middle/near audio and far/middle video data, respectively. [35] shows more recording details.

There are some variables that can be controlled during recording, for example, the TV/light can be turned on/off, etc. Moreover, speakers could be divided into several groups to discuss different topics, which results in higher overlap ratios. As shown in Table 2, we divide all combinations of variables into 4 categories based on noise and overlap.

In contrast to the challenge corpus, the updated corpus has the following features:

- During challenge, only far audio/video data in the evaluation set was released. In the updated corpus, all evaluation data has been released to support various research, such as middle-field speech, which could be used for lipreading task, etc.
- The time synchronizing problem for training and development sets was reported. We have double-checked all training/development data and corrected asynchronous samples.
- 10-hour new data has been added to the evaluation set for increasing the data diversity.

3. Framework

As shown in Eq.1, the recognition process can be formulated as the Bayesian decision problem, where \( w = [w_0, \ldots, w_n] \), \( X = [x_0, \ldots, x_{T-1}] \) and \( H \) denotes words, input features, and hypotheses, respectively. The Language Model (LM) \( p(w) \) is the probability of an \( n \)-characters sequence \( w \) and can be decomposed as Eq.2. With HMM, the Acoustic Model (AM) \( p(X|w) \) can be decomposed in the frame level as Eq.3 shown:

\[
\begin{align*}
&w^* = \arg \max_{w \in H} p(w|X) = \arg \max_{w \in H} p(X|w)p(w) \\
&p(w) = \prod_{t=0}^{n-1} p(w_t|w_{t-1}, \ldots, w_0) \\
&p(X|w) = \sum_y \prod_{t=1}^{T-1} a_{y_{t-1}y_t} \prod_{t=0}^{T-1} p(x_t|y_t) \\
&= \sum_y \prod_{t=1}^{T-1} a_{y_{t-1}y_t} \prod_{t=0}^{T-1} p(y_t|x_t)p(x_t)/p(y_t)
\end{align*}
\]

As shown in Fig.1, all systems are DNN-HMM systems as described above. The main difference lies in the X and p(X|w), which are described in Section 3.1 and Section 3.2 respectively. Section 3.3 introduces p(w), the lexicon and the metric.

3.1. Data Processing

Mel Filter Bank (FBANK) features are adopted as audio feature \( X_a \). A frontend consisting of weighted prediction error (WPE) dereverberation [36] and weighted delay-and-sum beamforming (Beamformitl) [37] is applied to the far-field 6-channel speech and the middle-field 2-channel speech before the FBANK feature extraction. For the near-field monaural speech, FBANK feature can be extracted directly from the raw audio waveform.

Lip Regions of Interest (RoIs) of the target speaker are cropped as visual feature \( X_v \). Because each frame contains only the target speaker in the middle-field video, an off-the-shelf face and lip detector is used to find face and lip, then a lip-centered window is cropped. All the speakers appear in each far-field video frame, we first detect all the faces and lips, then the face that best matches the mid-field video of the target speaker is selected.
Table 3: A overall comparison of CER among the DNN-HMM hybrid ASR/VSR/AVSR systems with far/middle/near audio and far/middle video over 4 recording configurations in Table 2.

<table>
<thead>
<tr>
<th>System</th>
<th>Audio</th>
<th>Video</th>
<th>CER (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>C1</td>
</tr>
<tr>
<td>A1</td>
<td>far</td>
<td>/</td>
<td>35.47</td>
</tr>
<tr>
<td>A2</td>
<td>middle</td>
<td>/</td>
<td>50.51</td>
</tr>
<tr>
<td>A3</td>
<td>near</td>
<td>/</td>
<td>25.71</td>
</tr>
<tr>
<td>V1</td>
<td>/</td>
<td>far</td>
<td>95.69</td>
</tr>
<tr>
<td>V2</td>
<td>/</td>
<td>middle</td>
<td>89.77</td>
</tr>
<tr>
<td>A1</td>
<td>far</td>
<td>/</td>
<td>30.99</td>
</tr>
<tr>
<td>A2</td>
<td>far</td>
<td>middle</td>
<td>45.57</td>
</tr>
<tr>
<td>A3</td>
<td>near</td>
<td>middle</td>
<td>22.64</td>
</tr>
</tbody>
</table>

Table 4: A comparison of $S$, $D$ and $I$ among A1 and A2 systems.

<table>
<thead>
<tr>
<th>System</th>
<th>$S$ (in %)</th>
<th>$D$ (in %)</th>
<th>$I$ (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TV</td>
<td>AV1</td>
<td>AV2</td>
</tr>
<tr>
<td></td>
<td>Off</td>
<td>On</td>
<td>Off</td>
</tr>
<tr>
<td>A1</td>
<td>48.16</td>
<td>35.05</td>
<td>17.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36.88</td>
<td>23.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.15</td>
</tr>
<tr>
<td>AV2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.2. DNN-HMM Hybrid Acoustic Model

The acoustic model consists of an embedding extractor and a followed sequence modeling model. The embedding extractor $f \in \{f_A, f_V, f_{AV}\}$ takes the audio feature $X_A$ or/and the visual feature $X_V$ as input to produce an embedding $E \in \{E_A, E_V, E_{AV}\}$. The sequence module is next employed to model the temporal dynamics. Finally, the posterior probability $p(y|X)$ is predicted by the ensuing full connection and Soft-Max layers. Multi-Scale Temporal Convolution Network (MS-TCN) [38] is adopted as the sequential modeling module in our systems. It can be formulated as:

\[
p(y|X) = \text{SoftMax}(FC(\text{MS-TCN}(E)))
\]

where the visual embedding extractor $f_V$ is the same as [39] which consists of a spatiotemporal convolution followed by an 18-layer ResNet. A spatiotemporal convolution consists of a convolution layer with 64 3D-kernels, a batch normalization, a ReLU activation and a spatiotemporal max-pooling layer. The audio embedding extractor $f_A$ has the similar structure as $f_V$. The 3D-kernels in spatiotemporal convolution and the 2D-kernels in ResNet-18 are replaced by 1D-kernels meanwhile the 3D-MaxPooling layer is dropped. The audio-visual embedding extractor $f_{AV}$ consists of a visual module $f_{AV}^V$ and an audio module $f_{AV}^A$ which have the same structure as $f_V$ and $f_A$, respectively. The concatenation is over the channel dimension and the frame mismatch between the audio and video is solved by repeating a video frame for several audio frames.

The Cross Entropy criterion $\mathcal{L}_{CE}$ between $p(y|X)$ and the true distribution of state $p^{GT}$ is calculated as:

\[
\mathcal{L}_{CE} = - \sum_{i=0}^{N_B-1} p_i^{GT} \log p(y_i|X)
\]

where $N_B$ is the minibatch size. $\mathcal{L}_{CE}$ is minimized by using Adam optimizer [40] for 100 epochs with an initial learning rate of 0.0003 and cosine scheduler [41]. The best model is selected by the highest classification accuracy on the development set $p^{GT}$ is generated with the GMM-HMM system. In ASR experiments, three GMM-HMM systems with far/middle/near audio are built. Far-field and near-field GMM-HMM systems are adopted for far-field and middle-field VSR, respectively. In AVSR experiments, the GMM-HMM system with the corresponding audio is adopted.

3.3. Decoding and Scoring

We use the DaCiDian dictionary as the pronunciation dictionary. A 3-gram language model is trained by the maximum entropy modeling method implemented in the SRILM toolkit.

We adopt Character Error Rate (CER) as the metric. It is represented with Eq. 9:

\[
CER = S + D + I
\]

where $S$, $D$, $I$ are substitution, deletion, insertion error rates, respectively. The lower the CER value (with 0 being a perfect score), the better the recognition performance. For speech overlap segments, we calculate all the substitution, deletion and insertion errors based on the recognition results and the ground truth for each speaker based on the oracle speaker diarization.

4. Experiments

Three ASR systems with far/middle/near audio and two VSR systems with far/middle video have been built as baselines for various research. For AVSR system, we considered three combinations: the most challenging combination of far audio and far video, the most ideal combination of near audio and middle video and the compromise of far audio and middle video. Far video to middle video could be achieved with the help of a zoom lens in real-world application. Table 3 shows a comparison of CER among all systems over 4 recording configurations in Table 2. Three ASR results show the performance degradation of ASR results from far-field channel distortion, ambient noises and speech overlap. Two VSR results show the performance of VSR is robust to noise and speech overlap but still very limited. The challenges mainly come from two aspects. Unfavorable lighting and variable head posture make it difficult for the far camera to capture lip changes, which is improved in the middle video. Chinese is a tone language and the tone is the use of pitch in language to distinguish lexical or grammatical meaning, which leads more words to look the same on the lip when pronounce.

4.1. Analysis of errors in noise condition

Based on the presence of noise, we divide the whole evaluation set into two categories and list the corresponding $S/D/I$ in Table 4. Table 4 reflects two trends. With the presence of TV noise, the $D$ of the A1 system rose sharply from 17.71 to 30.96, while neither $S$ nor $I$ rose much. AV2 system shows improvements over A1 across all evaluation metrics, specifically, there is a stable gain in $S$ whether noise’s presence or absence and larger improvement of $D$ in the noise condition. Due to the segmentation according to the oracle speaker diarization information during preprocessing, $I$ is very rare. Accordingly, we propose that video modality supplement acoustic information degraded by noise to reduce deletion errors.

In Fig.2, we show a noisy example selected from the evaluation set randomly and the comparison between the outputs of A1/V2/AV2 systems. In the second half of the far spectrum, the noise component almost completely drown out the target speech component, which causes four deletion errors in the results of the A1 system. As mentioned above, the challenges

from the environment and the tone language make the performance of V2 very limited. But the AV2 system utilizes the acoustic information in the change of lip shape to predict the last four characters correctly.

4.2. Analysis of errors in overlap condition

We think substitution errors are more related to speech overlap. So based on the number of speakers participating in the conversation, we divide the whole evaluation set into six categories and plot the corresponding $S$ in Fig.3. More speakers participating in the conversation means more overlapping speech. A sharp rise of $S$ follows to the number of speakers participating in the conversation is shown in Fig.3. AV2 system shows improvements over A1 across all six categories and larger improvement is observed in the category with more speakers.

In Fig.4, we show an overlapping example selected from the evaluation set randomly and the comparison between the outputs of A1,V2,AV2 systems. The target speech overlaps with the interfering speech almost completely, consequently, the output of A1 contained four substitution errors and a deletion error. V2 excels on the last character due to unambiguous pronunciation actions. AV2 corrected these five errors by utilizing the discriminative acoustic information in the change of lip shape.

4.3. Analysis on advanced techniques

Lastly, we explored some advanced techniques for the most challenging combinations of far-field audio and far-field video. Guided source separation (GSS) [42] is adopted as frontend and SpecAug [43] is applied during the acoustic model training, denoted as AV4 in 5. Further, we build three end-to-end AVSR systems which consist of a hybrid CTC/Attention based acoustic model [20] and a transformer-based language model, denoted as E2E1, E2E2 and E2E3 respectively. E2E1 and E2E2 are developed on a three-layer MS-TCN and a six-layer conformer based encoder backbone respectively using FBANK as audio input. E2E3 has a more flexible conformer-based encoder backbone with skip-connection and multiple fusion. The audio input is raw waveform and WavAug [44] is adopted during training. More details can be found in our source code.

As shown in Table 5, GSS and SpecAug yield significant performance improvements in the DNN-HMM hybrid AVSR system. The end-to-end AVSR system can achieve a comparable performance by simply changing. Besides, it is worth mentioning that many native end-to-end AVSR systems have been proposed on AVSR task in the MISP2021 challenge [35].

## 5. Conclusions

The updated AVSR corpus of MISP2021 Challenge is fully released. The corpus is the first distant multi-microphone conversational Chinese audio-visual corpus and the first large vocabulary continuous lip-reading Chinese corpus in the adverse home-tv scenario which promotes the research in the speech area, the computer vision area and cross-disciplinary area. Extensive experiments and error analysis show video supplement acoustic information to reduce deletion errors in noise condition and provide discriminative information in overlapping speech to reduce substitution errors. Finally, a set of experiments with advanced techniques provide the direction of the future work.

## 6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427 and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDC08050200.
7. References


